SHERLOCK: Experimental Evaluation of a Conversational Agent for Mobile Information Tasks

Alun Preece, William Webberley, Dave Braines, Erin G. Zaroukian, and Jonathan Z. Bakdash

***Abstract*—Controlled natural language (CNL) has great potential to support human–machine interaction (HMI) because it provides an information representation that is both human readable and machine processable. We investigated the effectiveness of a CNL-based conversational interface for HMI in a behavioral experiment called simple human experiment regarding locally observed collective knowledge (SHERLOCK). In SHERLOCK, individuals acted in groups to discover and report information to the machine using natural language (NL), which the machine then processed into CNL. The machine fused responses from different users to form a common operating picture, a dashboard showing the level of agreement for distinct information. To obtain information to add to this dashboard, users explored the real world in a simulated crowdsourced sensing scenario. This scenario represented a simplified controlled analog for tactical intelligence (i.e., direct intelligence of the environment), which is key for rapidly planning military, law enforcement, and emergency operations. Overall, despite close to zero training, 74% of the users inputted NL that was machine interpretable and addressed the assigned tasks. An experimental manipulation aimed to increase user–machine interaction, however, did not improve performance as hypothesized. Nevertheless, results indicate that the conversational interface may be effective in assisting humans with collection and fusion of information in a crowdsourcing context.**

***Index Terms*—Controlled natural language (CNL), conversational interface, human–computer collaboration (HCC), human–machine interaction (HMI), tactical intelligence.**

I. INTRODUCTION

CONTROLLED natural languages (CNLs) support human–machine interaction (HMI) or collaboration by providing an information representation that aims to be human readable and writable, while also being machine processable [1]. In the context of human–computer collaboration (HCC) [2], this means that CNLs provide a way to exchange information between human and software agents using a common mutually understandable language. While the design of CNLs for HCC involves tradeoffs between supporting robustness of machine processing and human friendliness,[1] CNLs are designed to be more human-friendly than traditional information and knowledge representations, offering the advantage of lower training overheads [3]. Simpler interfaces and ease of use have become key factors in the design of effective mobile applications ("apps") to support users performing tasks *in situ* with minimal training [4], where people are typically faced with a range of distractions caused by their environment, and software operation will often be secondary to their other activities. Also, the ability to quickly report, share, and fuse information is important in military and other safety-critical environments [5].

This paper reports behavioral research that assessed HMI with a CNL-based conversational agent, implemented as a mobile app. Individual users acting in groups collaboratively built a shared CNL knowledge base (KB) via a process of inputting natural language (NL) and confirming equivalent CNL suggested by the agent. App users explored the real world in a simulated crowdsourced sensing scenario, obtaining information (e.g., the color of Professor Plum's shirt) through observation and through interaction with actors in various roles. Because tactical intelligence (i.e., direct reports of events and situations in the environment) is key for planning military, law enforcement, and emergency operations in situations such as natural disaster relief or terrorist attacks, this scenario represented a simplified, but controlled, analog for tactical intelligence. The research design was motivated by a need to support users conducting information tasks *in situ*, for example, providing reports from the field on current events, seeking information relevant to their current situation, or instructing a range of "smart" devices—such as sensing systems or robots—to assist them. This motivation is consistent with the U.S. Department of Defence's Third Offset Strategy

[1]For example, to avoid ambiguity, CNLs allow only certain syntactic structures (e.g., "I shot an elephant in my pyjamas" might be expressed in a CNL as "I shot an elephant and I was in my pyjamas"/"I shot an elephant and the elephant was in my pyjamas"), but this can result in a language that is less natural to the user.

Manuscript received August 1, 2016; revised March 6, 2017; accepted April 9, 2017. Date of publication May 31, 2017; date of current version November 13, 2017. This work was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence and was accomplished under Agreement Numbers W911NF-06-3-0001 and W911NF-16-3-0001. The work of E. G. Zaroukian was supported by an appointment to the U.S. Army Research Laboratory Postdoctoral Fellowship Program administered by the Oak Ridge Associated Universities. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence, or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon. This paper was recommended by Associate Editor L. Chen. *(Corresponding author: Alun Preece.)*

A. Preece and W. Webberley are with the School of Computer Science and Informatics, Cardiff University, Cardiff, CF10 3XQ, U.K. (e-mail: PreeceAD@cardiff.ac.uk; WebberleyWM@cardiff.ac.uk).

D. Braines is with Emerging Technology Services, IBM United Kingdom Ltd., Winchester, SO21 2JN, U.K. (e-mail: dave_braines@uk.ibm.com).

E. G. Zaroukian and J. Z. Bakdash are with the Human Research and Engineering Directorate, U.S. Army Research Laboratory, Adelphi, MD 20783 USA (e-mail: erin.g.zaroukian.ctr@mail.mil; jonathan.z.bakdash.civ@mail.mil).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2017.2700625

for enhancing human capabilities through human–machine collaboration using artificial intelligence.[2]

Because the CNL representation of information for the conversational agent would be both human readable and machine processable, we hypothesize this would facilitate effective collaboration between humans and machine agents for the information tasks. The focus of our research is to assess if individuals, with minimal training, can effectively use the conversational agent to collaborate on information (gathering) tasks in mobile real-world settings.

Because of our focus on agent usability in a real-world environment, rather than a well-controlled laboratory environment, we use a quasi-experimental design, which has characteristics of both a study (no experimental manipulation) and an experiment [6]. Here, the nonexperimental aspect was agent usability and the experimental aspect was the manipulation of user interaction with the conversational agent. A quasi-experiment has tradeoffs compared to a well-controlled laboratory experiment: Greater generalizablility of results (to the real world), but increased susceptibility to confounding variables that may affect results (see [6] and Section V-D).

A. Motivation

We focus on assessing the usability of the conversational agent as a potential cognitive artifact, a tool that enhances human capabilities by externalizing aspects of cognition. Cognitive artifacts are defined as "artificial devices designed to maintain, display, or operate upon information in order to serve a representative function" [7]. The purpose of the conversational agent as a cognitive artifact was to enable users, who had minimal training in the CNL knowledge representation, to create a shared and dynamic KB for storing information. Once a KB exists, software agents can perform a variety of tasks to assist humans [5], [8][3]; the goal of this work was not to examine those kinds of task but merely to test that users can create a KB that is machine processable.

Framing the use of CNL in conversational-style interactions between human and machine is partly inspired by the recent resurgence of interest in NL interfaces. Widespread access to commercial products such as Apple's Siri,[4] Amazon's Alexa,[5] Google Now,[6] and IBM's Watson[7] provides users with a familiar context in which to conduct information tasks *in situ*. In contrast to widely used commercial products, our conversational interface focused on building a KB (e.g., North Road is blocked by heavy snow) rather than merely querying an existing one (e.g., What is the weather like today?).

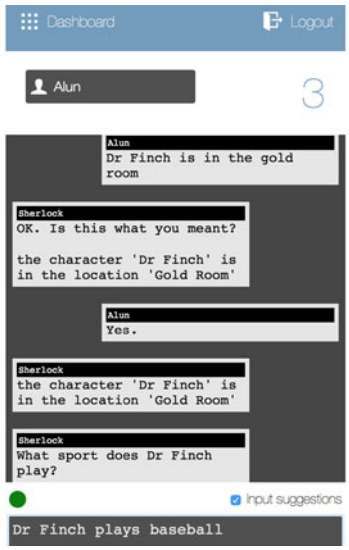

Fig. 1. Screenshot of the conversational agent, illustrating the conversational protocol.

Interaction with the conversational agent is controlled by a protocol based on linguistic speech act theory [9]. The effects of speech acts on the KB are persistent and affect subsequent interactions, for example, if a user tells the agent something, and some user subsequently asks about the same thing, the agent's response should reflect what it has previously been told. This principle, that the state of the world is modified by an exchange of messages, has underpinned much of the work in software agent communication languages [10]. The conversational interface in the experiment presented here operated in two conditions, corresponding to the range of speech acts supported: a condition in which the user and software agent could ask each other questions, and one in which they could not. We hypothesized that the condition with ask speech acts would facilitate greater general usability with the conversational interface.

In keeping with users' common expectations regarding mobile conversational agents, and to minimize training overheads, it is unrealistic to expect human users to communicate with software agents using CNL exclusively. An important feature of the experiment reported here is that users are permitted to interact with the conversational agent in NL. Before acting upon any NL input, the agent will attempt to interpret the NL by generating a piece of CNL that it will ask the user to *confirm*. Confirmation by the user will permit the agent to act upon the received message. For example, if the message is a query, then the agent will try to provide a response; if it is a piece of new information, then the agent will attempt to integrate it into its current KB and share it with other users and agents. An example of this can be seen in a screenshot of the CNL conversational agent in Fig. 1, where the user Alun tells the agent Sherlock a piece of information in NL. Sherlock then translates this into CNL and asks Alun to confirm, Alun confirms it, and this information is added to the KB. (The agent's restatement of the confirmed CNL is intended as feedback to the user that the statement was added to the KB.)

[2]https://web.archive.org/web/20170112191920/https://www.defense.gov/News/Article/Article/628154/work-human-machine-teaming-represents-defense-technology-future

[3]As a simple example of the kind of information aggregation task agents can do once the KB is populated, users were provided with a visualization of the shared KB in the form of a dashboard.

[4]https://web.archive.org/web/20160628202009/http://www.apple.com/ios/siri/

[5]http://web.archive.org/web/20170208020449/https://developer.amazon.com/alexa

[6]https://web.archive.org/web/20170110125710/https://en.wikipedia.org/wiki/Google_Now

[7]https://web.archive.org/web/20160628201902/http://www.ibm.com/watson/

B. Hypotheses

To summarize, the two main hypotheses are as follows.

1) *Overall usability (nonexperimental):* The conversational agent would have high usability as an effective cognitive artifact. Usability was operationalized as performance [11]: the number of user-inputted NL messages that were both machine interpretable and confirmed by the user.
2) *Agent interaction capability (experimental):* The conversational agent would be more usable with a broader range of speech acts compared to no broader range of speech acts.

This paper is organized as follows: Section II reviews background and related work in HCC. Section III describes the CNL-based approach used in the experimental evaluation. Section IV describes the design of the experiment. Section V presents our results. Section VI reflects on the outcomes of the research and points to future work.

II. Related Work

This research focuses on the use of CNLs for HCC, where humans and machines work together. The use of CNL facilitates machine assistance for users, but the focus is not on human automation in the sense of reducing human input or control [12]. Similarly, in the context of sensor and information fusion, the focus of the research is not on technical aspects of collecting and processing data and information (e.g., topics such as sensor capabilities, network bandwidth, and algorithms for sensor and information fusion) but on the ability of humans to understand machines to make informed decisions [13], [14].

Most prior research on social sensing (data and information derived from humans, such as geolocation, search engine terms, and social media like Facebook or Twitter) to infer situations and events has been observational [15]. Consequently, inferences made using social sensing can be quite wrong [16]. Methods do exist to validate social sensing data [15] (e.g., the veracity of social media statements is typically assessed with probabilistic uncertainty bounds using computational approaches), but validation is fairly uncommon because ground truth is rarely obtainable. The scenario-based research design presented in this paper provides a ground truth and so avoids inferential pitfalls of not having correct answers to evaluate against.

Despite the resurgence of interest in intelligent language-understanding systems, open problems include how to imbue machines with more natural conversational behaviors including turn-taking and user interruptions [17], and how to operate effectively beyond static domains [18] to reduce problems of brittleness common in these kinds of systems. Mass-market intelligent agents such as Siri and Google Now remain essentially confined to simple ask–tell interactions rather than flowing conversations. User familiarity with these modes of use led us for this experiment to confine ourselves in the main to ask-tell-style interactions, with emphasis on reduction of ambiguity through confirmatory interactions.

The conversational approach is one type of HCC in which humans and intelligent systems work together with a common goal [2]. There is a growing body of HCC work in relation to collaborative situation awareness and intelligence analysis. Analysts are increasingly well versed in modern collaboration environments and social media, and systems are emerging that seek to combine the benefits of these approaches with existing software tools and processes for structuring and supporting tactical intelligence analysis. A recent example of this in [19] seeks to enable analysts to identify the decision-relevant data scattered among databases and the mental models of other personnel by employing familiar social media-style collaboration techniques. There is some evidence to indicate that not only is it useful to collaborate within the same analyst team/group, but, when collaboration is extended to the crowd and mediated by an intelligent software agent, the outcome of the intelligence analysis can be greatly improved [20]. The authors propose a web-based application to collate imagery of a particular location from media sources and provide an operator with real-time situation awareness. Such approaches are promising, showing that a richly collaborative environment—social, HCC, or both—can be a blessing if machines can help in sorting, filtering, and managing large amounts of information. However, the same approaches can be a curse if the volume of information is simply increased.

The broader context and application domain for the current research is facilitating intelligence processes for the rapid cycles of information collection, interpretation, and decision making. Management of sensing assets for intelligence, surveillance, and reconnaissance traditionally follows a well-known cycle, referred to in the UK as DCPD: direction, collection, processing, and dissemination [21]. The U.S. variant of the DCPD cycle called TCPED—tasking, collection, processing, exploitation, and dissemination—refers to direction as "tasking" and divides the processing step into two parts, "processing" and "exploitation," where the former is essentially preprocessing to put data into a usable form, and the latter involves putting the information into the context of a particular decision. In the context of the experiment reported here, the conversational agent has the ability to direct/task humans (as sensors) via a mobile app, by asking them questions. Information collection is done via the humans telling the agent answers. Processing/exploitation is carried out by the agent, which assembles a picture of the entire situation. This picture is then disseminated to the humans via the mobile app.

III. Approach

As mentioned in the introduction, the design of CNLs typically involves tradeoffs in terms of the complexity of the language from both the human and machine perspectives; simpler linguistic forms are easier for machines to process robustly, but these can seem awkward and unnatural for humans and are limited in terms of what can be expressed. Conversely, more natural linguistic forms for humans can lead to ambiguity and loss of robustness in machine processing.

The research presented here uses a particular form of CNL developed through the Network and Information Sciences International Technology Alliance (ITA)[8] called ITA Controlled

[8]http://nis-ita.org

English (CE), designed for low linguistic complexity and to eliminate ambiguity in expressions [22]. The choice of this particular CNL was determined by such research being part of a larger body of work using ITA CE to support decision making and collaborative tasks. The language will be referred to simply as CE in this paper.

A. Controlled English

The CE language specification includes linguistic constructs for defining conceptual models (ontologies), instances (facts), and rules, although the latter are outside the scope of this paper. For illustration, a sample CE model definition is shown as follows:

```
conceptualise a ~ character ~ C that
  is a sherlock thing and
  is a locatable thing and
  has the color C as ~ shirt color ~.
conceptualise the character C
  ~ works for ~ the organization O and
  ~ eats ~ the fruit F and
  ~ likes ~ the hobby H.
```

These two CE `conceptualise` sentences define a new concept in a CE model (ontology). New model terms—concepts, properties, and relationships—are introduced between the tilde (`~`) symbols. The first sentence introduces the new concept `character`, defining it as being a child of the parent concepts `sherlock thing` and `locatable thing`. The first sentence also contains a property definition for the `character` concept (denoted by the `has` keyword, `shirt color`, and the type of the property value, `color`). The second sentence expands the definition of `character` by adding relationship definitions and the types of the related things, for example, a `character works for` an `organization` and `eats` a kind of `fruit`. Being a child of the concepts `sherlock thing` and `locatable thing` means that `character` also inherits any properties and relationships defined on its parent concepts, for example, the concept `locatable thing` has a relationship `is in` with instances of the concept `location`.

Instances (facts) are defined in CE using the following syntax. The following example shows an instance of the concept `character`, as defined above.

```
there is a character named 'Prof Plum'
  that is in the location 'N215' and
  eats the fruit 'banana' and
  has the color 'white' as shirt color.
```

This instance is named `Prof Plum` and, due to being also an instance of `locatable thing`, has an `is in` relationship with an instance of `location`, called `N215`. The instance `Prof Plum` also has an `eats` relationship with an instance of the concept `fruit`, `banana`, and a value for its `shirt color` property, `white`.

These examples are drawn from the domain of the research described below in Section IV. While this domain is deliberately simplistic, CE has also been used extensively in real-world applications including mission planning [23], intelligence analysis [24], and coalition knowledge management [25]. Moreover, modeling in CE is intended to be flexible, supporting the creation of extensible models with whatever concepts, properties, and relationships are needed. CE models can be extended at runtime, though this feature was not explored in the current research.

B. Conversational Protocol

While more human-friendly than traditional information and knowledge representations, the above examples demonstrate that the design of CE favors robust machine processing over naturalness. For example, the need for syntactic marker phrases such as `there is a` in instance definitions, and the specification of type information in property and relationship expressions, can make the resulting sentences seem cumbersome and may reduce the human user's speed and accuracy of comprehension. One approach to addressing this issue is via tool support, for example, the provision of syntax-directed editing and autocomplete. Another approach is to allow humans to use NL and to provide software that mediates between NL and CE.

Full details of a protocol to support conversations that flow between NL and CE are given in [26]. The experiment reported in Section IV focuses on two main types of interaction:

1) *confirm* interactions where the initiator issues an NL message, which the receiver attempts to re-express in CE, and seeks confirmation from the issuer that the CE version is an acceptable interpretation of their message[9];
2) *ask–tell* interactions where the initiating agent issues a query (*ask*) and the receiver responds in some way it deems to be appropriate (usually by answering the query with a piece of information, that is, a *tell*).

An example interaction illustrating the use of this protocol is discussed in the next subsection.

The protocol is necessary to control the flow of conversations not only to make explicit the expectations on what the receiver of a message of a particular kind should do, as is traditional in speech act-based agent communication protocols [10], but also to avoid any potential for ambiguity in whether a particular piece of text is NL or CE. The *confirm* interactions mark the boundaries between NL and CE in the human–machine conversation.

C. Prototype Conversational Agent

The prototype conversational agent app was implemented to test the effectiveness of CE for machine-assisted performance of information tasks in a mobile setting. A screenshot is shown in Section I in Fig. 1. The conversation between the user (in this case, "Alun") and the agent ("Sherlock") is displayed in the style of a conventional smartphone text "chat" thread. The users type their NL input into the panel at the bottom. The users' messages then appear shifted to the right of the main display, with the agent's responses appearing on the left.

The first three messages comprise a *confirm* interaction. The initial message from the user ("Dr Finch is in the gold room") is in NL. The agent uses a relatively simple "bag of words"

[9]The protocol allows for any interaction to be initiated by human or machine; however, in this case, the initiator is usually a human.

algorithm (representing user input as a multi-set of word occurrences, disregarding grammar) [5] to interpret this in terms of a CE model of the world. Its attempt is shown in the second message ("the character 'Dr Finch' is in the location 'Gold Room'"), which the user then confirms is acceptable.[10] Following a positive confirmation, the agent then repeats the confirmed CE form as acknowledgment—a CE *tell* message)—and adds it to its KB. If the user rejects the agent's interpretation, then nothing is added to the agent's KB and the user can try again by rephrasing the message. Depending on the agent's configuration, it may also share this information with other agents (via a CE *tell*). In this configuration of the agent, either the user or the agent can ask questions of the other party; the example shows the software agent asking the user a question ("What sport does Dr Finch play?") in the bottom most message (initiating an *ask–tell* interaction), with the user shown to be entering his or her response in the input panel ("Dr Finch plays baseball").

The agent is configurable into multiple variants to test alternative experimental conditions. For example, some variants have the ability to support *ask* messages from the user and/or the agent; some support syntax-directed autocomplete (shown by the tick above the user input panel); some can operate either in offline mode or online, while others assume a network connection (in the screenshot, the online status of the agent is shown by the green marker above the input panel). The purpose of the number to the right of the username on the display, and the dashboard button, is explained in Section IV.

The prototype agent is implemented as a Web app in JavaScript to run in Web browsers on a variety of devices, including iOS and Android smartphones and tablets, as well as laptop and desktop computers.

IV. User Evaluation

The primary intent of the user evaluation was to gather evidence for whether the CNL-based approach to HCC on information tasks described in the previous section can be used effectively by users, with low training overheads. To maintain ecological validity in supporting small teams at the tactical edge, it is important that:

1) participants should be assigned to carry out information tasks *in situ*;
2) successful use of the CNL-based approach should allow users to gain measurable assistance from software agents from their perspective;
3) there should be some element of human–human collaboration as well as HCC [27].

The design of the experiment was motivated by a desire to emulate aspects of tactical intelligence tasks typically carried out by military, law enforcement, and other individuals on patrol in field settings, particularly where cooperation is needed between members of multiple partners in a coalition [8]. In the design, the information tasks were simplified to allow participation without any specific tactical intelligence training and to ease some aspects of the natural language processing (NLP) performed by the conversational agent (since NLP was not the focus of the research).

In the experiment, the conversational agent was designed to assist the participants in their tasks by collecting disparate pieces of user-reported information into a shared KB. To motivate participants to provide information, they received an individual score with one point for each piece of *confirmed* information that was relevant to the assigned tasks; this score is shown on the app in Fig. 1 to the right of the displayed username. We refer to each piece of task-relevant confirmed information as an *assertion*; an individual's score is a count of their assertions. Participants were also able to view a visualization of the shared KB in the form of a dashboard (accessible via the button at the top left of the app in the figure). Note that, while it was possible for participants to enter information into the KB that was not relevant to the assigned tasks, they would not earn any points for such input.

Groups of participants were assigned the same tasks and given different variants of the conversational agent and supporting infrastructure. To address Hypothesis 1, which involves no experimental manipulation, we measured usability, operationalized by performance and quantified by *total participant assertions* (i.e., the successful performance of the operation of adding information to the KB). To address Hypothesis 2, the dependent variable in this experiment again was usability, with the level of agent interaction capability (*confirm only* versus *confirm and ask–tell*) as the independent variable.

Both the experiment and the conversational agent were called Simple Human Experiment Regarding Locally Observed Collective Knowledge (SHERLOCK), a name chosen to give participants a sense that their tasks may involve elements of detection.

A. Participants

Participants were drawn from a sample of convenience: they were second- and third-year UK undergraduate students studying human–computer interaction and knowledge management at Cardiff University. To fit into the students' timetable, the experiment was run three times over two days, with each student attending one session. Individuals were randomly assigned to equal-sized groups in advance but some opted not to attend their assigned session. This resulted in three groups (A, B, and C), consisting of 19, 9, and 11 members, respectively ($N = 39$). While there was no explicit requirement for equal-sized groups, we discuss issues arising from this nonhomogeneous group size in Section V-D.

The experiment was run immediately following a 50-min lecture on the general principles and applications of CNLs and a brief demonstration of the use of the conversational agent. Because our primary interest was in the usability of this agent with little to no prior training, participants were given no opportunity to practice using the agent before participating in the experiment.

[10]While this final confirm takes extra time, voice communications have standard confirmation terms (e.g., roger, wilco, copy). It remains an open question whether this would be appropriate for HCC/HMI, and this is worth considering in future work.

B. Design and Hypotheses

Recall that Hypothesis 1 (overall usability) was nonexperimental, whereas Hypothesis 2 was an experimental manipulation of the conversational interface. The experimental design was a single factor with two levels between participants. In the confirm condition, the group members were given access to a limited version of the conversational agent supporting only *confirm* interactions, that is, they were able to submit messages to the agent in NL and *confirm* (or not) the CE of what the agent understood. In the ask–tell condition, each group member was given access to a fuller version of the conversational agent supporting question answering (*ask–tell* interactions) initiated by both human and machine, in addition to *confirm* interactions. For overall usability, our hypothesis was that participants would be able to utilize the conversational agent to make assertions to the shared KB. As for the role of conversational protocol (level of agent interaction capability), our hypothesis was that ask–tell condition participants would submit more assertions as the additional features would enable them to 1) become more aware of how the agent processed the CNL and thus able to communicate with it more effectively and 2) be directed by the agent to provide specific information.

Because of our focus on agent usability, we opted not to have a control group with no agent. We were not aiming to show that using the CNL agent is better (or worse) than a manual process, but rather that CNL is a usable medium for crowdsourced knowledge collection and processing.

Group A participants were assigned to the confirm condition, while Group B and Group C participants were assigned to the ask–tell condition. Members of Groups B and C were shown that they could ask questions of the agent in addition to confirming CE. In our regression analysis (see Table III in Section V), group and condition were analyzed because it was possible scoring differences may be due to either factor.

Postexperiment, subjective usability was assessed by asking participants to complete the system usability scale (SUS)[11] and a questionnaire to determine

1) their prior experience with NL search engines (e.g., Apple's Siri or Google's NL search);
2) their prior awareness of the game "Cluedo"/"Clue";
3) the extent to which they shared information with others (outside of the use of the app);
4) the extent to which they received information from others (outside of using the app).

C. SHERLOCK Game

The game was designed to be played in a complex of university buildings with which the participants were expected to be generally familiar. Participants were given a sheet of paper listing 54 questions. Note the questions had ground truth, a single correct answer, unlike most prior social sensing research. Questions were designed to encourage participants to physically visit locations around the building complex to discover the answers. Participants were encouraged to use their own mobile device (typically a smartphone or tablet) to access the conversational agent and answer the questions *in situ*; they were also permitted to use a device (tablet or desktop computer) provided in various key locations. Users were randomly assigned to groups to minimize any difference in performance due to variations in device functionality or user familiarity with the devices.

TABLE I
SUMMARY OF PARTICIPANT GROUPS, CONDITIONS, AND INSTRUCTIONS

Condition	Confirm	Ask-tell
Participants	Group A $N = 19$	Group B, Group C $N = 9 \quad N = 11$
Agent Capabilities	Confirm	Ask-tell, confirm
Instructions	You have 2 tasks. This is a *crowd-sourcing* game — work as individuals within your group. Your group is in competition with the other group for the highest group score. Task 1: Use the SHERLOCK agent to submit as many answers as you can to the questions on your sheet. You'll get one point for each new answer you submit that SHERLOCK understands. Task 2: Use the SHERLOCK agent to report things you see in (or near) any of the locations mentioned in the questions. You'll get one point for each report you submit that SHERLOCK understands. You'll get points even for mundane objects like an iPad, a printer, or a mouse.	

The participants were given a unique randomly assigned username and instructed to use this to identify themselves to the conversational agent. Participants were given 30 min to complete as much of the task as possible. The farthest locations were separated by no more than a 5-min walk.

Participants were instructed as shown in Table I. A prize was offered to the highest scoring participant across the groups, that is, the participant with the most assertions.

The two tasks were designed to simulate closed- and open-ended tactical intelligence activities respectively, aimed at collecting "who/what/where/why" information.[12] Most of the "who/what/where/why" elements centered on six characters portrayed by human actors, one in each of six rooms referenced in the participants' questions. The majority of the 54 questions in Task 1 referred to these six characters[13]—Reverend Green, Colonel Mustard, Sergeant Peacock, Professor Plum, Captain Scarlet, and Doctor White—and their distinct attributes, e.g., their shirt color, a particular kind of fruit in their possession, their hobby, their employer, or their emotional state (e.g., sad). Participants needed to physically visit and enter the six locations to discover the character, their attributes, and any other objects in the room. Some attributes such as shirt color and fruit could be determined by observation; others such as hobbies or employers could only be determined by questioning the character. Access to each location was restricted to a single participant at a time, so participants could not confer on the room's contents while *in situ*. In case this might result in participants queuing to access a busy location, the two tasks were designed to give them plenty of alternative things to do instead of waiting. Example questions included:

[11] http://hell.meiert.org/core/pdf/sus.pdf

[12] The environment was considered static during each experiment run, so "when" information was out of scope, an environment that changes over the time of the experiment will be a feature of future work.

[13] Their names were based on the classic "Clue"/"Cluedo" game.

TABLE II
FEATURES SCORING INDIVIDUAL POINTS

	"Synthetic" features	"Natural" features
Task 1 (closed)	Features of the six characters, including their location, shirt color, fruit, hobby, employer, emotional state	Locations of specified real-world objects (artworks and artifacts)
Task 2 (open-ended)	Locations of "anomaly objects" (balloons, gorilla, dinosaur)	Locations of "mundane objects" such as office equipment and furniture

```
What character eats bananas?
What character is wearing a red shirt?
What is the hobby of Professor Plum?
Who does Captain Scarlet work for?
Why is Colonel Mustard sad?
Where is the lemon?
```

To make Task 1 more challenging (mimicking real-world tactical intelligence tasks), a minority of the 54 questions asked participants to report the locations of distinctive real-world objects, such as artworks and artifacts, distributed in public areas around the building. Example questions of this kind included:

```
Where in Queen's Buildings is the
racing car?
Where in Queen's Buildings is the
aeroplane wing?
Where in Queen's Buildings is the
painting of Edmund Hann?
Where in Queen's Buildings is the
Penydarren train plaque?
```

Anticipating that some of these public locations might be hard for participants to describe to the agent, guidance was given as to how to express them, for example, "South building stairs," "Central building second floor," "North building lobby."

Users could gain points for their assertions in Task 2 by reporting the locations of a variety of "mundane objects" including pieces of office equipment and furniture. In addition, six "anomaly objects" were placed in or near the six character-containing locations, intended to mimic suspicious objects in real-world tactical intelligence tasks. None of these objects were things conventionally found in the environment. These were: a large heart-shaped balloon, a Mickey Mouse balloon, a 6-foot-tall inflatable dinosaur, a 4-foot-wide inflatable soccer ball, a large pink balloon in the shape of the numeral "6," and a toy gorilla. None of these were referred to explicitly in the set of 54 questions for Task 1. We were interested in the extent to which participants in the different conditions would focus their efforts on Task 1 versus Task 2, and the extent to which they would report the "anomaly objects" in particular. Table II summarizes the kinds of features and objects participants could gain points for observing in each of the two tasks. Note that, while the questions were designed to encourage participants to visit locations in the buildings, in some cases, it was possible that a participant might have known the location of one of the real-world objects, or might have been told an answer by another participant. Neither of these cases would affect the goal of assessing the usability of the conversational agent, however, only the degree to which participants roamed the buildings.

Fig. 2. Shared dashboard summarizing a group's performance on Task 1.

The app dashboard visualizing the status of the shared KB is shown in Fig. 2. The 54 squares correspond to the 54 questions in Task 1 (being open-ended, there was no simple way to visualize participants' progress on Task 2 beyond incrementing their individual assertion score). The color of the grid square corresponding to each question indicated the state of collected information relevant to that question: *gray* (the starting state) indicated that no information had yet been submitted by any participant that answered the question; *amber* indicated that some (consistent) information had been obtained, but not enough to provide a "settled" answer to the question; *green* indicated that enough information had been obtained to provide a "settled" answer to the question; and *red* meant that the information obtained indicated multiple conflicting answers to the question. Heuristics were used to differentiate between amber, red, and green states, aimed at encouraging participants to provide more information to turn as many squares green as possible.

Participants were told that their group would be awarded ten points for each question they collectively made green on the dashboard.[14]

D. Summary of SHERLOCK Design for Tactical Intelligence Relevance

Rather than aiming to fully recreate the real-world task, simulation-based training and assessment should incorporate psychologically relevant aspects from the real-world task and the environment [28]. Consequently, the simplified task presented here incorporates several key psychological characteristics of a real-world tactical intelligence task:

1) *Time pressure:* Participants had a finite amount of time to "complete" the task.

[14]Recall that Group A was roughly double the size of Groups B or C. Groups B or C then may have had greater difficulty in moving a square from the amber ("unsettled") state than Group A did, and this had the potential to affect morale and performance.

2) *Too much information and high uncertainty:* The task was intentionally designed to be impossible to complete. There was ambiguity in locations and interactions with the characters.

Moreover, the live network environment was a realistic feature, exemplified by a server drop-out for Group A (see Section V); although this was an unintended aspect, network connectivity issues are common in military environments [27].

One could argue the experiment was confounded and the task were oversimplified relative to real-world operational environments. However, if we were unable to establish effective use of the conversational interface with a "simplified" experimental design, it is extremely unlikely it would be effective under the much more challenging conditions for military, law enforcement, and others in safety critical real-world environments.

V. Results and Discussion

The primary results were that the conversational agent had high usability, supporting the first hypothesis that it would act as an effective cognitive artifact. Results from the SUS provide converging evidence. However, the second hypothesis—that added speech act capabilities would increase usability—was not supported; there was no meaningful difference between the confirm and ask–tell conditions in terms of individual scores or the number of reported anomalies. Exploratory analyses of the secondary results provide insights into the primary results. Finally, we discuss the tradeoffs and limitations with the experimental design.

Time duration and reproducible research: Twenty-two minutes into the 30-min experiment, data collection for Group A was affected by a server drop-out.[15] To account for the server drop-out, presented results were conservatively analyzed using the first 22 min from all groups, unless noted otherwise. Critically, results were comparable regardless whether or not the drop was considered. Data and full results, including analyses with and without the time cutoff, are available online: http://osf.io/pz529. All results are fully reproducible.

A. Primary Results

Hypothesis 1: As explained in Section IV, usability was assessed as objective performance [11], operationalized as assertions (i.e., the successful performance of the operation of adding information to the KB). Recall that individual participants received one point for each assertion, that is, a *confirm* interaction that ended in a submission of a piece of CE to the agent, either in an attempt to answer one of the 54 questions (Task 1) or to report an object in the environment (Task 2). An awarded point indicated that users had made themselves understood to the agent and had made a contribution to the collective KB relevant to the assigned tasks.

A histogram of participants' total assertions from the three groups is shown in Fig. 3: 29 out of 39 (74%) users had one point or more; 2 out of 39 (5%) had only one point; 27 out of 39 (69%) had more than one point. Fig. 4 shows individual participants' cumulative assertion counts during the periods of each group. The trajectories generally show steady upward progression, with a few cases of significant growth in the latter part of the run.[16]

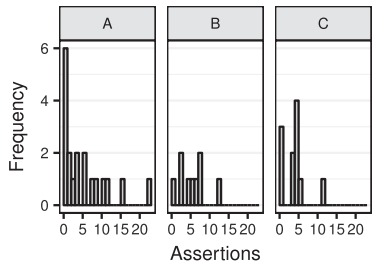

Fig. 3. Histogram of participants' total assertions from the three groups.

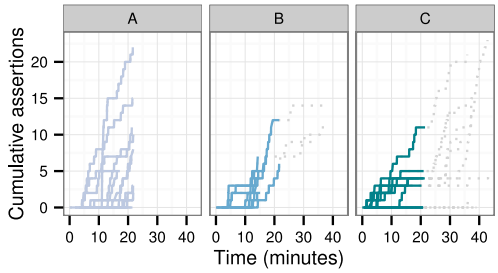

Fig. 4. Individual participants' cumulative assertions for each group. Assertions made after 22 min are shown dotted in gray.

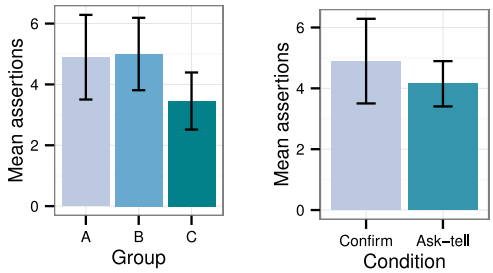

Fig. 5. Mean assertions: (left) per group (Groups A, B, and C) and (right) per condition (confirm condition, ask–tell condition), bars represent one standard error of the mean.

These results provide evidence that the conversational approach can be effectively used with close to zero training: a sizeable majority of users were able to become productive in using the agent, that is, add to the KB, in a relatively short period, operating *in situ* while attending to multiple simultaneous information tasks.

Hypothesis 2: The mean assertions for each group (Groups A, B, and C) are shown on the left of Fig. 5. The right of the figure shows the mean assertions for each experimental condition; recall that participants in the confirm condition used an agent equipped only with *confirm* conversational capability, while participants in the ask–tell condition used an agent

[15]The drop-out occurred 21 min and 47 s into the experiment; for brevity, we will refer to this as 22 min.

[16]The data for Group A are truncated due to the server drop-out near the end of the period; the data for Group C indicate that some participants were able to continue submitting data beyond the 30-min cutoff.

TABLE III
REGRESSION ANALYSIS BY GROUP AND CONDITION

By group: initial $\theta = 0.92$ (dispersion parameter)				
Coefficients	Estimate	Standard Error	z-value	p-value
Intercept	1.59	0.26	6.09	< 0.001
Group B	0.02	0.46	0.05	0.96
Group C	−0.35	0.44	−0.79	0.43
By condition: initial $\theta = 0.91$ (dispersion parameter)				
Coefficients	Estimate	Standard Error	z-value	p-value
Intercept	1.59	0.26	6.05	< 0.001
Ask–tell capability	−0.17	0.37	−0.45	0.66

equipped with both *ask–tell* and *confirm* conversational capabilities. Error bars represent one standard error of the mean. These results show that the assertion totals did not differ significantly by group or by condition.

In performing a regression analysis on the assertion counts, counts were overdispersed, that is, the Poisson distribution assumption of rate = variance was clearly violated [29]. Regressions, therefore, used a negative binomial distribution, which permits the rate and variance to differ. Another regression method using a quasi-Poisson distribution, also recommended for analyzing overdispersed data, produced similar results. A summary of the analysis by both group and condition is shown in Table III. The p-values indicate that the difference in participants' performance both by group and by condition were not significant.[17]

Participants reported positive satisfaction based on scores from the SUS. SUS results did not vary widely between the three groups, with means in the high 60 s indicating a good degree of usability. Detailed results from the accompanying postexperiment questionnaire are not presented here for brevity. The response rate was 57.5%.

In terms of the group scoring rule—ten points for each "settled" question (i.e., green on the dashboard)—Groups A and B each scored 80, while Group C scored 160. These are the end-of-run scores seen by the participants on their dashboards, used as the basis to reveal the highest group score, and do not reflect the server drop-out experienced by Group A.

Anomaly objects were reported by a small number of participants (eight out of 39 participants reported one or more such objects), so results are only presented descriptively (see Fig. 6). In most of the cases, the agent was able to properly interpret these anomaly reports (i.e., inputs mentioning an anomaly object and a location). However, in a few cases (five out of 14 messages), we subsequently identified messages that were clearly intended to be reports of an anomaly object but where the agent had failed to interpret them satisfactorily; we counted these "failed" reports along with the properly interpreted ones because we were interested in the extent to which users apparently noticed the anomaly objects. More anomaly objects were reported with *ask–tell* than *confirm*. However, given the limited data, no strong conclusions can be drawn.

[17]Data analyses were performed using R [30] and the *lme4* package [31].

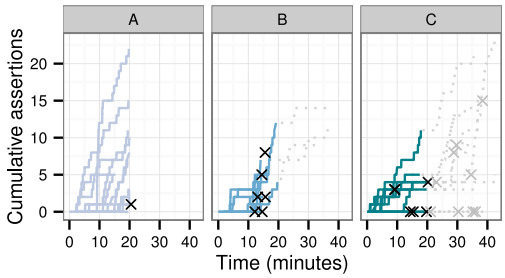

Fig. 6. Anomaly messages over time, denoted by "X." Messages are superimposed over individual participants' cumulative assertions. Note overlapping "X"s were jittered to improve visibility.

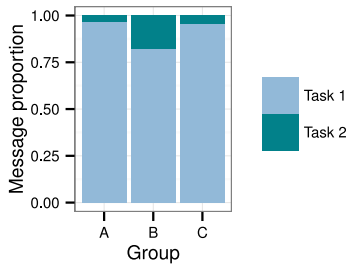

Fig. 7. Proportion of user-submitted messages addressing Task 1 and Task 2.

1) Discussion: Most of the users (74%) were successful in using the conversational agent, with accelerating assertions over time indicating fast learning (see Fig. 4). We interpret this as strong evidence that the conversational agent was a highly usable cognitive artifact and is supported by converging results from the usability questionnaire [11]. There were few reports of anomaly objects (see Fig. 6). Without more data, we can only infer that differences in speech act support by the agent had a minimal impact on anomaly reporting in this experiment.

Because there were no meaningful differences among groups or conditions, in terms of usability, the experiment provides no evidence that the enhanced conversational capabilities of the agent equipped with speech acts to support question-answering either led to improved productivity in generating assertions or any difference in participants' attention to Task 1 versus Task 2. The absence of a clear difference is explored next in the secondary results, as are reasons for why users might have concentrated on Task 1—addressing the 54 (dashboard) questions—rather than Task 2—objects in the environment. Possible reasons why the *ask–tell* capability went largely unused are also discussed.

B. Secondary Results

In the secondary results, graphical summaries of the messages by group are presented in Figs. 7–9.

1) Fig. 7: Message proportions for Task 1 versus 2.
2) Fig. 8: Message counts: submitted, interpreted, and confirmed.
3) Fig. 9: Message proportions for a question versus a statement.

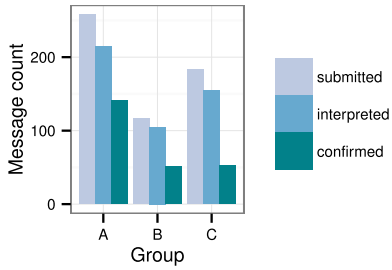

Fig. 8. Counts of messages submitted by the user, interpreted by the agent, and confirmed by the user.

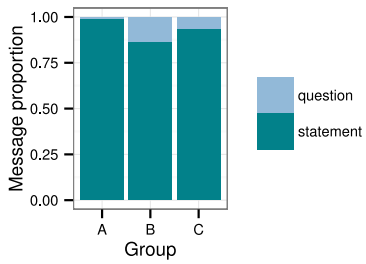

Fig. 9. Proportion of user-submitted statements and questions.

Overall, a high proportion of messages (93%) addressed Task 1 not Task 2. Fig. 7 shows the proportions of submitted messages that related to the various parallel activities groups were tasked to: Tasks 1 and 2 refer to the instructions given to the participants to 1) try to answer the 54 questions and 2) report on objects in the environment not explicitly referred to in the 54 questions.

Across groups, 85% of submitted messages were interpreted by the agent (see Fig. 8). The proportion of interpreted messages confirmed by users was notably low (52%) (see Fig. 8); the most common reasons for this are discussed below.

In Fig. 8, examples of "submitted," "interpreted," and "confirmed" are as follows. In the example interaction shown in Fig. 1 in Section III:

1) "Submitted": The user's initial NL input, "Dr Finch is in the gold room";
2) "Interpreted": The agent's translation of this in to "the character 'Dr Finch' is in the location 'Gold Room"' indicates that this message was "interpreted";
3) "Confirmed": The user's subsequent confirmation of the CE would count this as a "confirmed" message.

The vast majority of messages (95%) were statements rather than questions.

Fig. 9 classifies submissions into whether they were statements of fact, intended to impart information (e.g., to answer one of the 54 questions or to report an object in the environment), or whether they were "questions" addressed to the agent. Recalling that the Group A members, in the confirm condition, were given an agent that was incapable of responding to questions, it is interesting to see that a few attempts were made to query it regardless. More interestingly, it is evident that members of the ask–tell condition groups did not make extensive use of their agent's query capability, particularly in Group C.

1) Discussion: Overall, participants focused their efforts overwhelmingly on Task 1 (answering the 54 questions), with only Group B making non-minimal efforts to address Task 2 (reporting objects in the environment). There are several complementary explanations for this emphasis on Task 1: First, Task 1 was more "visible" to participants due to the question sheet and the prominence of the dashboard (which showed progress on Task 1 only). Second, Task 2 may have been seen as having lower value than Task 1 because it did not contribute to the group score (which was based on the number of "green" questions on the dashboard). Third, Task 1 was closed, whereas Task 2 was open, potentially leading participants to focus on a task they felt they could complete.

There are also several complementary reasons to explain why members of the ask–tell condition groups did not make extensive use of their agent's query capability. First, the scope of Task 1 was large enough that participants could continue to make progress in "lighting up the dashboard" without needing to use their agent's query capability. Second, the query capability had the most value for helping resolve conflicted questions. With so many questions remaining in an uncertain state throughout the experiment, it seems participants likely perceived that they had "more than enough to do." Consequently, it is feasible that increasing the duration of each run and/or reducing the number of questions would have led to greater use of the agent's query capability because greater efforts would have been allocated to resolving conflicted questions.

Other exploratory results indicated that the majority of submitted messages were interpreted by the agent, although a lower number were confirmed. In many cases, this was due to the agent's failure to properly interpret users' attempts to describe the locations of real-world objects. In other cases, it was due to the agent not recognizing named entities that we had not anticipated users mentioning, e.g., the real names of the actors playing the characters. In some cases, however, it appears that many users simply forgot to confirm the agent's perfectly accurate interpretations of their input.

C. Observations

Actors playing the six characters were able to observe that there was some degree of organisation within groups to the extent that particular users tended to visit the same locations together and take turns to enter the designated rooms, or to send one individual in, who would then gather information and report it directly to the others. This was borne out in patterns of message submission. Other expected submission patterns included repeated attempts to be understood by the agent, in some cases mimicking CNL styles apparently in an effort for their input to be properly processed.

As noted above, participants encountered particular difficulties in communicating the locations of real-world and anomaly objects leading to some frustration in both Tasks 1 and 2. In some cases, users were unable to work out how to provide input that resulted in CNL (messages "submitted but not interpreted" in terms of Fig. 8); in other cases, the agent generated CNL that the user opted not to confirm ("interpreted but not confirmed").

These cases caused many of the real-world object questions to remain gray or amber on the dashboard. In other cases, users provided conflicting descriptions of a location, resulting in red squares on the dashboard. The following are examples of a participant struggling to describe a location (expected answer: South building ground floor):

```
The aeroplane wing is in the basement
The aeroplane wing is in the workshop
The aeroplane wing is in p980
```

D. Tradeoffs and Limitations

This research has several tradeoffs and limitations, many because the SHERLOCK experiment used an open, uncertain, real-world environment, and a nonexperimental design for usability. While we demonstrated that the agent is usable, a methodological limitation of the nonexperimental design is no experimental comparison to another system or a control condition. Additional shortcomings include: different group sizes, informal information sharing among users (users directly communicating with each other), server drop out, ambiguity in location and character interactions, knowledge of the environment, and use of personal devices. Although some of these limitations were intentional, others were not. Minimizing potential confounds, especially unintended ones, in future research will increase internal validity.

Moreover, we used a sample of convenience, and this sample was small at the group level but more than sufficient at the individual level because of repeated measures over time. The total number of input messages with the conversational interfaces was 558, a mean of 14.31 $(558/39)$ per participant.

Few anomalies were reported. However, by definition, anomalies are rare or unusual, which was our intention. Making them easier to find and/or providing explicit incentives for reporting them would substantially change the task. Nevertheless, this is a limitation of the design.

As noted above, some users exhibited confusion in naming the real-world locations in Task 1 or referenced entities outside the scope of the game. This would not have been an issue had we chosen to use a list of predetermined, selectable entities, and locations in the agent. However, specifying this information would have turned the experiment into more of a matching task than using a conversational agent. Furthermore, it is not always possible to provide, in advance, all objects and locations of potential interest in the real world. These tradeoffs may account for the (self-reported) degree of frustration among a small number of the participants. Enhancements to the agent for deciphering location names would help mitigate this issue. Finally, minor user interface changes to confirmation (e.g., a reminder to confirm and/or autoconfirmation after a small amount of time) are highly likely to substantially increase confirmations.

In addition, a few participants confused the characters with the actors playing them, using their real-world names and affiliations rather than the character ones. This could be addressed with stronger clarification in the instructions or by replacing human actors with synthetic characters (e.g., cartoon scenes depicted on posters).

Last, the time and effort to run the experiment (seven experimenters, with substantial setup time) in the real world was nontrivial. There was a tradeoff by conducting the research in a real-world environment rather than a more controlled laboratory environment. It is possible that participant performance would have been much higher with a well-controlled "clean" laboratory experiment, but the greater internal validity would likely have come at expense to external validity. While the server drop-out for group A was an unintended confound, it is a common occurrence in the real world and did not meaningfully impact the results.

VI. Conclusion and Future Work

Results from the SHERLOCK experiment provide evidence that untrained users were able to become productive in a short time using the conversational agent to provide information on a situation. The experiment was unable to confirm whether the more sophisticated question-answering capability is helpful as this capability was used only to a very limited extent by participants, most likely due to time pressures of the task. Also, the design of the task provided little incentive for users to use the capability. We will address this in the design of future experiments.

Feedback from the agent to the user in the experiment was confined to the use of CE in *confirm* interactions. An area for future experimentation would be to compare this with more "natural" forms of feedback (e.g., textual or spoken NL, or graphical feedback as was explored briefly in [26]). Further experiments will examine wider styles of conversation, general usability of the CE form of CNL, ability to quickly model or extend a model in a domain, multiuser conversations, and potentially also conversations with multiple different NLs—particularly important in coalition operations.

In terms of the practical functionality of the conversational agent, a key future objective is to extend the agent so that it is able to acquire input from more sources, e.g., audio/image/video input from the mobile device, metadata such as the device type/model, spatial and temporal data, and potentially even cues as to its user's emotional state. For example, the system could be extended to support crowdsourcing via social media by having the conversational agent operating behind a Twitter account so that it could, for example, use Twitter to collect information (either from public accounts or by asking) or disseminate that information by retweeting.

Finally, the use of CNL to support machine–machine as well as human–machine conversations is highly applicable to the Internet of Things (IoT) context [32]. IoT is networked "smart" physical devices with software and sensors that are typically both automated and user controlled. In particular, CNL enables a common representation for machine–machine interactions that is also amenable to human understanding. We plan to conduct future experiments in the IoT context.

References

[1] T. Kuhn, "A survey and classification of controlled natural languages," *Comput. Linguist.*, vol. 40, pp. 121–170, 2014.

[2] L. Terveen, "Overview of human-computer collaboration," *Knowl.-Based Syst.*, vol. 8, no. 2, pp. 67–81, 1995.

[3] R. Schwitter, K. Kaljurand, A. Cregan, C. Dolbear, and G. Hart, "A comparison of three controlled natural languages for OWL 1.1," in *Proc. 4th OWL Experiences Directions Workshop*, 2008.
[4] R. Harrison, D. Flood, and D. Duce, "Usability of mobile applications: Literature review and rationale for a new usability model," *J. Interact. Sci.*, vol. 1, 2013, Art. no. 1.
[5] A. Preece, C. Gwilliams, C. Parizas, D. Pizzocaro, J. Z. Bakdash, and D. Braines, "Conversational sensing," *Proc. SPIE*, vol. 9122, 2014, Art. no. 912201.
[6] W. Shadish, T. Cook, and D. Campbell, *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston, MA, USA: Houghton Mifflin, 2002.
[7] D. A. Norman, "Cognitive artifacts," in *Designing Interaction*, J. M. Carroll, Ed. Cambridge, U.K.: Cambridge Univ. Press, 1991, pp. 17–38.
[8] A. Preece, D. Pizzocaro, D. Braines, D. Mott, G. de Mel, and T. Pham, "Integrating hard and soft information sources for D2D using controlled natural language," in *Proc. 15th Int. Conf. Inf. Fusion*, 2012, pp. 1330–1337.
[9] J. Austin and J. Urmson, *How to Do Things With Words*. Cambridge, MA, USA: Harvard Univ. Press, 1975.
[10] Y. Labrou and T. Finin, "Semantics and conversations for an agent communication language," in *Readings in Agents*, M. N. Huhns and M. P. Singh, Eds. San Mateo, CA, USA: Morgan Kaufmann, 1998, pp. 235–242.
[11] J. Nielsen, *Usability Engineering*. Buffalo, NY, USA: AP Professional, 1994.
[12] M. Cummings, "Man versus machine or man + machine?" *IEEE Intell. Syst.*, vol. 29, no. 5, pp. 62–69, Sep./Oct. 2014.
[13] J. Z. Bakdash, D. Pizzocaro, and A. Preece, "Human factors in intelligence, surveillance, and reconnaissance: Gaps for soldiers and technology recommendations," in *Proc. IEEE Military Commun. Conf.*, 2013, pp. 1900–1905.
[14] E. Blasch, "Level 5 (user refinement) issues supporting information fusion management," in *Proc. 9th Int. Conf. Inf. Fusion*, 2009, pp. 1–8.
[15] D. Wang, T. Abdelzaher, and L. Kaplan, *Social Sensing: Building Reliable Systems on Unreliable Data*. San Mateo, CA, USA: Morgan Kaufmann, 2015.
[16] D. Lazer, R. K. G. King, and A. Vespignani, "The parable of Google Flu: Traps in big data analysis," *Science*, vol. 343, pp. 1203–1205, 2014.
[17] H. Hastie *et al.*, "Demonstration of the Parlance system: A data-driven, incremental, spoken dialogue system for interactive search," in *Proc. SIGDIAL Conf.*, 2013, pp. 154–156.
[18] M. Gašić *et al.*, "POMDP-based dialogue manager adaptation to extended domains," in *Proc. SIGDIAL Conf.*, 2013, pp. 214–222.
[19] A. Wollocko, M. Farry, and R. Stark, "Supporting tactical intelligence using collaborative environments and social networking," *Proc. SPIE*, vol. 8758, 2013, Art. no. 87580E.
[20] R. Brantingham and A. Hossain, "Crowded: A crowd-sourced perspective of events as they happen," *Proc. SPIE*, vol. 8758, 2013, Art. no. 87580D.
[21] G. Pearson and T. Pham, "The challenge of sensor information processing and delivery within network and information science research," *Proc. SPIE*, vol. 6981, 2008, Art. no. 698105.
[22] D. Mott, "Summary of ITA Controlled English," 2010. [Online]. Available: http://nis-ita.org/science-library/paper/doc-1411a
[23] J. Patel, M. Dorneich, D. Mott, A. Bahrami, and C. Giammanco, "Improving coalition planning by making plans alive," *IEEE Intell. Syst.*, vol. 28, no. 1, pp. 17–25, Jan./Feb. 2013.
[24] D. Mott, D. R. Shemanski, C. Giammanco, and D. Braines, "Collaborative human-machine analysis using a controlled natural language," *Proc. SPIE*, vol. 9499, 2015, Art. no. 94990J.
[25] C. Giammanco *et al.*, "Knowledge management for coalition information sharing at the network edge," *IEEE Intell. Syst.*, vol. 28, no. 1, pp. 26–33, Jan./Feb. 2013.
[26] A. Preece, D. Braines, D. Pizzocaro, and C. Parizas, "Human-machine conversations to support multi-agency missions," *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 18, no. 1, pp. 75–84, 2014.
[27] M. S. Vassiliou, D. S. Alberts, and J. R. Agre, *C2 Re-envisioned: The Future of the Enterprise*. Boca Raton, FL, USA: CRC Press, 2014.
[28] F. A. Drews and J. Z. Bakdash, "Simulation training in health care," *Rev. Human Factors Ergon.*, vol. 8, no. 1, pp. 191–234, 2013.
[29] A. Zeileis, C. Kleiber, and S. Jackman, "Regression models for count data in R," Dept. Statist. Math., WU Vienna Univ. Econ. Bus., Vienna, Austria, Res. Report Ser., Tech. Rep. 53, 2007.
[30] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Found. Statist. Computing, 2014. [Online]. Available: http://www.R-project.org/
[31] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *J. Statist. Softw.*, vol. 67, no. 1, pp. 1–48, 2015.
[32] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of things (IoT): A vision, architectural elements, and future directions," *Future Gener. Comput. Syst.*, vol. 29, no. 7, pp. 1645–1660, 2013.

Alun Preece received the Ph.D. degree in computer science from the University of Wales, Swansea, U.K., in 1989.

He is the Co-Director of the Crime and Security Research Institute, Cardiff University, Cardiff, U.K. From 2011 to 2014, he was an Academic Technical Area Leader for the Network and Information Sciences International Technology Alliance (NIS ITA). His research interests include decision support, distributed knowledge-based systems, and human–computer collaboration.

William Webberley received the B.Sc. (Hons.) degree in computer science in 2010 and the Ph.D. degree in solving the "unfiltered feed" problem experienced in online social networks in 2015, both from Cardiff University, Cardiff, U.K.

Since then, he has been a Researcher as part of the NIS ITA team in the School of Computer Science and Informatics at Cardiff University, focused on distributed and decentralized agent–agent and human–agent systems for knowledge base building and task assignment.

Dave Braines received the B.Sc. (Hons.) degree in computer science from the University of Portsmouth, Portsmouth, U.K., in 1993.

He is the CTO for the Emerging Technology team at IBM United Kingdom Ltd., Winchester, U.K. Since September 2016, he has been an Industry Technical Area Leader for the recently formed International Technology Alliance in Distributed Analytics and Information Science (DAIS ITA) research program. His research interests include effective human/machine hybrid teams and the technologies needed to underpin them.

Mr. Braines is a Fellow of the British Computer Society.

Erin G. Zaroukian received the Ph.D. degree in cognitive science from Johns Hopkins University, Baltimore, MD, USA, in 2013.

She is currently a Postdoctoral Fellow with the Human Research and Engineering Directorate, U.S. Army Research Laboratory, Adelphi, MD. Her research interests include natural language semantics and human–machine interaction.

Jonathan Z. Bakdash received the Ph.D. degree in psychology from the University of Virginia, Charlottesville, VA, USA, in 2010.

He is a Research Psychologist with the Human Research and Engineering Directorate, U.S. Army Research Laboratory, Adelphi, MD, USA. His research interests include human decision making, human–machine interaction, and cybersecurity.